Dynamic Operator Overload Estimation during Supervisory Control of Multiple UAVs

Leonard A. Breslow[1], Daniel Gartenberg[2], J. Malcolm McCurry[3], and J. Gregory Trafton[1]
Naval Research Laboratory, [2]George Mason University, [3]ITT Exelis

Abstract-- **Crandall et al. and Cummings & Mitchell introduced fan-out as a measure of the maximum number of robots a single human operator can supervise in a given single-human-multiple-robot system, based on the time constraints imposed by limitations of the robots and of the supervisor, e.g., limitations in attention. Adapting their work, we introduced a dynamic model of operator overload that predicts failures in supervisory control in real time, based on fluctuations in time constraints and in the supervisor's allocation of attention, assessed by eye fixations. Operator overload was assessed by damage incurred by vehicles when they traversed hazard areas. The model generalized well to different tasks. We then incorporated the model into the system where it predicted in real-time when an operator would fail to prevent vehicle damage and alerted the operator to the threat at those times. These model-based adaptive cues reduced the damage rate by one half relative to a control condition.**

I. INTRODUCTION

As robots become cheaper and more autonomous, there is an opportunity to enable one human supervisor to control multiple robots simultaneously. Yet increasing the number of robots that are controlled can hinder operator performance in time-critical supervisory control tasks by increasing operator workload, thereby impacting the operator's attentional resources. Understanding the factors that determine the effectiveness of the overall human-robot system, including factors that affect the cognitive state of the operator, can contribute to the development of adaptive automation that can improve operator performance

One measure of the number of robots a single operator can supervise at one time is Crandall et al.'s fan-out equation [1] . Fan-out predicts the maximum number of robots that can be monitored by taking into account how much time can pass before a robot needs to be acted on ("neglect time") in comparison to the length of time required for an operator to interact with a robot needing attention until it no longer requires attention ("interaction time") [1]. More precisely, neglect time (NT) is the amount of time a robot can be ignored by the operator before its performance drops below some predetermined level, and interaction time (IT) is the amount of time required for the operator to interact with the robot in order to restore the robot's performance to the predetermined acceptable level. The more autonomous the robot or UAV, the longer its NT and consequently, the higher the fan-out, i.e., the number of robots a single operator can control. Similarly, the less the interaction time (IT), the higher the fan-out.

Cummings and Mitchell [2] extended this work on fan-out by adding a stronger emphasis on the perceptual and cognitive processes of the operator. Specifically, they included in their fan-out computation *wait time* variables, including delays in allocating attention to a vehicle requiring help (WTAA) and delays due to task queuing (WTQ), when there are delays due to allocating time among several vehicles that require attention at the same time. These wait times constitute time demands in addition to IT that limit fan-out.

Fan-out is a useful global assessment of a particular task reflecting the demands the task places upon the operator, thereby facilitating cognitive engineering design and improving training. We will explore whether the dynamic variability of performance during the course of a particular task can be predicted by the same, or

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Dynamic Operator Overload Estimation during Supervisory Control of Multiple UAVs	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory ,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
IEEE Transactions on Human-Machine Systems, vol. 44, no. 1, pp. 30-40, 2014.

14. ABSTRACT
Crandall et al. and Cummings & Mitchell introduced fan-out as a measure of the maximum number of robots a single human operator can supervise in a given single-human-multiple-robot system, based on the time constraints imposed by limitations of the robots and of the supervisor, e.g., limitations in attention. Adapting their work, we introduced a dynamic model of operator overload that predicts failures in supervisory control in real time, based on fluctuations in time constraints and in the supervisor???s allocation of attention, assessed by eye fixations. Operator overload was assessed by damage incurred by vehicles when they traversed hazard areas. The model generalized well to different tasks. We then incorporated the model into the system where it predicted in real-time when an operator would fail to prevent vehicle damage and alerted the operator to the threat at those times. These model-based adaptive cues reduced the damage rate by one half relative to a control condition.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**21**	

similar, factors that predict fan-out for a task, or human-robot system, as a whole. Presumably, even where the operator is supervising no more robots than prescribed by the fan-out equation, there will be moments when events converge to make him/her vulnerable to temporary overloading and, therefore, to error. We shall refer to transitory overload of this sort as dynamic *operator overload.*

We hypothesize that measures of fan-out can be adapted for predicting dynamic operator overload and so provide a basis for preventing operator errors of commission or omission. When the predicted likelihood of error due to overloading rises, adaptive cues can be introduced to reduce wait time and prevent errors.

II. From Fan-out to Dynamic Operator Overload

A. Limits of Supervisory Control: Fan-Out

Crandall et al. [1] proposed that the maximum number of robots that could be controlled by a single human operator, or fan-out (FO), could be computed as:

$$FO = NT/IT + 1 \quad \textbf{(1)}$$

This equation defines fan-out as the maximum number of vehicles an operator can interact with (i.e., the number of IT intervals) during the NT of another vehicle that does not currently require interaction. The "+1" in the equation accounts for the latter, neglected vehicle.

Cummings and Mitchell [2] extended this fan-out equation to include the human factor wait times WTAA and WTQ. These are combined with IT in the denominator of the ratio:

$$FO = (NT / (IT + WTAA + WTQ)) + 1 \quad \textbf{(2)}$$

where each of the terms (i.e. NT, IT, WTAA and WTQ) are *sums* over the course of a session. As these sums enter into a ratio, the result is similar to the computation based on mean values. While fan-out is a global measure of operator capacity on a task, the amount of operator overload within arbitrary time intervals during the course of a task can be expected to fluctuate as the fan-out variables drift from their respective mean values. Such fluctuations alter the probability of overload-induced errors over the course of a supervisory control task. Our goal is to instantiate these fan-out variables in a model designed to predict when operator load increases enough to cause operators to become overloaded and as a consequence make errors.

In subsequent work, Crandall and Cummings implemented stochastic models of operator-vehicle interactions based on traces of interaction sessions [3, 4]. These models predict the operator's selection of a vehicle to handle and predict vehicle state, based on observed sequences of vehicle states and selections. While this approach was successful in predicting operator performance across task variations, it did not analyze cognitive factors or within-task performance variation, the main foci of the work to be reported here.

B. Predicting operator overload in a supervisory control task

In the current research, we attempted to predict when an operator supervising multiple UAVs (unmanned aerial vehicles) will become overloaded. The simulated control system we used automatically assigned each UAV to a target and determined its initial trajectory towards that target. In addition, there were threats, or hazard areas, that would cause a UAV to be damaged if not avoided. Participants could add waypoints to the trajectory or re-assign a UAV to a different target, in an effort to prevent damage to a vehicle. Once a UAV arrived at its target, the operator directed it in delivering its payload on the target.

The episodes of interest were path-intersects threat (PIT) events, which start from the moment a vehicle enters on a collision course with a threat and ends either at the point in time when the vehicle traverses the threat area, incurring damage, or else at the point the vehicle changes course away from the threat due to the operator's evasive actions. It is clear when a UAV will traverse a threat area, as a vehicle's trajectory to its target is displayed by a line, which intersects the threat in these cases. However, participants are not specifically alerted to the threat. We assumed that an

operator was overloaded when he/she allowed a UAV to incur damage by traversing a threat. We attempted to predict when an operator will fail to prevent a vehicle from taking damage by incorporating variables similar to those in the fan-out equation within a model designed to predict dynamic operator overload.

As a matter of terminology we will define the *focal vehicle* of a PIT event to be the vehicle that is on a threat trajectory during that event. Likewise, the threat and target towards which the focal vehicle is heading will be referred to as the focal threat and focal target, respectively, of the PIT event, and collectively, together with the focal vehicle, as the *focal objects*. Vehicles, targets, and hazards other than the focal objects will be referred to as *non-focal objects*. Multiple PIT events may overlap in time, producing one of the main challenges of multiple-vehicle supervision. WTQ represents the amount of time devoted to subtasks related to non-focal vehicles.

The requirements of the particular problem under investigation motivated a minor change in the fan-out equation, in which the fan-out variable neglect time (NT) was replaced by the variable available time (AT). AT is the time interval from the start of the PIT event to the expected time of impact with the threat. During the AT interval, the operator needs to take care of the focal vehicle that is on the threat trajectory, as well as other vehicles requiring attention during that interval, if possible. The number of vehicles that an operator can handle during the AT intervals, including the focal vehicle, is the fan-out. Thus, for the purposes of the present research, we modified fan-out Equation 2 to

$$FO = AT / (IT + WTAA + WTQ) \quad \textbf{(3)}$$

where neglect time (NT) is replaced by available time (AT). Also, the result is not incremented by 1 as it was in Equation 2 because the 1-increment represented a vehicle that can be neglected and there is no vehicle that can necessarily be neglected during a PIT event.

Dynamic operator overload was assessed within each PIT event as the occurrence of damage to the focal vehicle. The variables considered for our model predicting damage in a PIT event were operationalized as follows:

1. *Wait Time Attention Allocation (WTAA):* the amount of time it took to recognize that the focal UAV requires attention. This was operationalized as the duration from the start of a PIT event until the relevant threat was first looked at..

2. *Task Queuing:* represents the allocation of attention to non-focal vehicles. Two alternative variables were considered:
    a. *Wait Time Queue (WTQ):* As in the fan-out model, WTQ represents the amount of time spent on manual actions on non-focal objects.

    b. *Wait Queue Fixations (WQF)*: the number of eye fixations on non-focal objects.

3. *Available Time (AT):* the interval from when a vehicle enters on a collision course with a threat (i.e., the start of a PIT event) until it will make contact with the threat if successful evasive action is not taken. This is the amount of time available to the operator to recognize and remedy the threat problem, and often includes excess time during which other vehicles can be maintained or monitored.

Note that the fan-out variable IT is not included in our dynamic model of operator overload. In the present context, IT is the time spent on actions resulting in the successful avoidance of damage during a PIT event. Since we are trying to predict the occurrence of damage on a per-event basis, IT is trivially related to damage. Thus, IT does not contribute to our understanding of the processes involved in the occurrence or prevention of damage.

The predictor WQF replaced WTQ, in our model, in part, because it was based on eye fixations, rather than manual actions. Similarly, the predictor WTAA is measured by eye fixations. Eye fixations are a more comprehensive measure of cognitive focus than manual actions, since eye fixations accompany cognitive processes, such as attention allocation,

situation assessment, and planning, which can occur with or without concurrent manual actions. In addition, as we shall see, the predictive model based on the eye fixation variable, WQF, was somewhat superior to the model based on manual actions, WTQ.

An eye-tracker was used in this research to record operator's fixations on a computer screen. Eye-trackers are able to measure where an operator is looking (called a fixation) and how long they look at something (called the fixation duration)[5, 6]. Several eye movement measures have been shown to be indicators of cognitive processing [5-7]. We used eye fixations as a measure of operator attention allocation. While it is possible to look at a stimulus without attending to it [8], eye movements have been found to correlate with attentional shifts [9-11]. As a covert shift of attention seemingly precedes an eye movement to the target of a saccade, eye movements can serve as a direct measure of attention [10]. In addition, the examination of eye movements has been used to predict procedural errors in a manner similar to the present research [12, 13].

Our predictive model of damage in PIT events was computed using logistic regression analysis. Logistic regression computes a multiple linear regression model with a dichotomous outcome variable; a more detailed description can be found in [14]. The dichotomous outcome variable in our analysis of PIT events was the occurrence/avoidance of damage to the focal vehicle. Unlike other classifiers, logistic regression allows one to determine whether or not each of the predictor variables had a statistically significant impact on the overall success of the model, in addition to assessing the model as a whole. Logistic regression has been used in predictive models of procedural errors in previous research [12, 13].

C. Overview of the Paper

Our predictive model was created and evaluated over five experiments. In Experiment 1, we generated the model. Experiment 2 was used to replicate and validate the model by assessing its application to an experimental condition identical to that in Experiment 1. Experiments 3 and 4 assessed the generalizability of the model to different task conditions that were, respectively, relatively easier or more difficult than Experiment 1. Experiment 5 assessed whether the model could predict operator overload in real time by generating cues to warn participants of threats. Specifically, we incorporated the model into the supervisory control simulation to provide real-time cues of upcoming threats that the model predicted would cause damage and compared performance on this system to performance on the system without cues.

III. EXPERIMENT 1. BASELINE FOR MODEL GENERATION

To examine the cognitive processes underlying operator attention and time allocation in a supervisory control task, data were collected from a complex dynamic supervisory control simulation. In the simulation, the participant controlled five semi-autonomous, homogenous (UAVs. The high-level goal of the simulation was to direct UAVs to specific targets on a map and visually identify key items at the target site in order to deliver the payload on those items. As participants performed the simulation, eye movement and mouse data were collected.

A critical component to successfully completing the simulation was to prevent UAVs from passing over threat areas, which periodically changed position on the map in an unpredictable manner. If a UAV "hit" (i.e., traversed) a threat area, the UAV took damage and could become incapacitated. Each time a UAV's path intersected a threat area, the operator had to take an explicit action to divert the UAV and prevent damage.

A. Method

1) Participants

Thirty-five George Mason University undergraduate students, 14 male and 21 female, participated for extra credit in a psychology course. All participants had normal or corrected-to-normal vision. Participants were asked to rate how often they played video games on a scale of one (never), two (sometimes), or three (a lot). The average amount of video game play was 1.9 (SD=.6).

Figure 1. The RESCHU supervisory control simulation, consisting of 3 panels: A. payload panel, B. map panel, C. status panel.

2) Simulation Description

The supervisory control task, originally designed as the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) [15, 16], was modified to only include homogenous UAVs moving at a constant speed of 5.2 pixels per second. The interface of the supervisory control simulation, shown in Figure 1, has three main sections: the map panel, the status panel, and the payload panel. The map panel (Figure 1, panel B) displays UAVs (blue half ovals), targets (red diamonds) towards which UAVs are moving, and threats (yellow circles) which should be avoided by UAVs. The status panel (Figure 1, panel C) shows the status of the UAVs and includes information on vehicle damage, time until the vehicle reaches a waypoint or target, and time remaining in the simulation. The payload panel (Figure 1, panel A) is used for acquiring a payload target through a visual acquisition task (described below) which is performed after a UAV reaches a target and the target is engaged by the operator.

The operator's high level goal in the simulation is to monitor UAVs as they proceed to specific target areas in the map panel, to engage the targets, and to perform a payload visual acquisition task once the UAV has engaged the target. The payload visual acquisition task requires the participant to search for an object (i.e. the target) in the randomly selected image within the payload panel and to mouse-click the target, thereby causing the payload to be delivered to the target. During the visual acquisition payload task, the operator cannot make any actions on the map panel of the interface, even though the vehicles continued to progress towards their targets on the map panel and new PIT events could begin during this interval.

Throughout the session, five UAVs moved along straight-line trajectories towards an automatically-assigned target. There were also eighteen threat areas. Every four seconds, one of the eighteen threats was randomly selected to change its position, with the constraint that the threats could not appear closer than 3° of visual angle (about 50 pixels) away from any UAV. If the UAV passed through a threat, it incurred damage. Damage was indicated as a bar in the status panel. The appearance of targets and threats on the simulation map was randomized with the constraint that targets and threats could be no closer than 3° of visual angle from each other.

If a UAV passed through a threat area, the UAV incurred damage. A UAV eventually became incapacitated ("dead") if it incurred enough damage. To avoid a threat area, the operator could perform two possible actions. First, the operator could direct the UAV to a different target, thereby changing the path of the UAV and avoiding the threat area. Second, the operator could add, delete or move waypoints on the UAV's trajectory to divert the UAV from the path to the threat, without changing the final destination target.

There were always 7 targets present on the map. At the start of the simulation the UAVs were randomly assigned to different targets; thus, the UAVs might not be directed towards the optimal target. After a target was engaged and the visual acquisition payload task was complete, the UAV was randomly assigned to a new currently-unassigned target which again might not be optimal. The simulation was a complex task with multiple events happening in parallel. More than one UAV could be waiting at their respective targets for engagement at the same time and more than one UAV could be on a path intersecting a threat area at a time.

When performing the simulation, participants were scored on their performance, both the number of targets that were engaged correctly and the amount of damage incurred by vehicles. However, no overall score was presented. Participants were instructed to engage as many targets as possible and to prevent damage to vehicles as much as possible. In the version of RESCHU we implemented the damage to a UAV inflicted by a threat was severe and could permanently incapacitate the UAV. Participants were provided ongoing feedback in the status panel on the amount of damage incurred by vehicles and on the number of incapacitated vehicles.

3) Design and Procedure

Prior to the start of the experiment, participants completed an interactive tutorial that explained all aspects of the simulation. During the tutorial, participants learned the objective of the simulation, how to control the UAVs (assigning targets, changing targets, assigning waypoints), and how to engage a target and complete the visual acquisition task in the payload panel. Participants were also warned of the dangers of threats and were instructed on how to avoid threats. The tutorial lasted approximately ten minutes. After completing the tutorial, the experimenter went over all of the controls with the participant to ensure that the participant understood the task.

After completing the practice session, participants were seated approximately 66 cm from the computer monitor and were calibrated on the eye tracker. Participants were again instructed to engage as many targets as possible and prevent as much damage as possible. Participants then began the simulation session, which lasted for 10 minutes. When the simulation session ended, participants received feedback on how many vehicles they engaged and total vehicle damage. Then, participants were re-calibrated and were run in a second 10-minute session with identical procedures to the first session. The data from both sessions were combined in the analyses to be presented.

4) Measures

The data from the supervisory control task were segmented into PIT events. Keystroke and mouse data were collected for each participant. Eye tracking data were collected using an SMI RED eye tracker operating at 250 Hz. A fixation was defined using the dispersion method based on a minimum of 15 eye samples within 60 ms and within 50 pixels (approximately 3º of visual angle) of each other, calculated in Euclidian distance. Three areas of interest were defined: UAVs, threats, and targets. Other fixations on the map panel and fixations on the payload panel were not analyzed. The eye tracker and the RESCHU simulation were synchronized, such that the simulation sent the eye tracker an update of its state each time its state was updated, i.e., every 500 ms.

We calibrated a participant on the eye tracker until each eye had a visual angle of less than one degree. After 10 unsuccessful attempts to calibrate a participant, the participant was not included in the data analysis. Calibration took less than 5 minutes.

B. Results

For the 35 participants in the experiment, there were a total of 1,999 PIT events, 216 (10.8%) of which ended in damage to UAVs. Mean duration of PIT events was 14,916.2 ms (SD=16.33). The other main action performed by participants, payload delivery (visual acquisition), had a mean duration of 4,800 ms (SD=400).

1) Developing a Logistic Regression Model

To create a logistic regression model of the PIT events, the outcomes of damage and no damage were coded as a binary outcome variable for each PIT event. The four predictor variables of interest (WTAA, WQF, AT) were recorded for each of the PIT events. WTAA and AT were recorded in ms. WQF was an integer representing the quantity of non-focal fixations during a PIT event. Equation (4) represents our dynamic overload model as a logistic regression equation predicting damage outcomes of PIT events:

$$\text{Predicted Logit of Damage} = 2.17 + (.00007 * \text{WTAA}) + (.11 * \text{WQF}) - (.00027 * \text{AT}) \quad \textbf{(4)}$$

The output of a logistic regression model is called a logit; its use in prediction will be explained later. This model was computed based on the final values for each PIT event. Thus, the model is dynamic across PIT events, but not within PIT events. In the final experiment, we will examine whether the model is useful for dynamic prediction within PIT events.

The overall logistic regression model was significant, X^2 (3) = 240.68, p<.0001. The log odds of damage was significantly related to each of the three predictors (p<.0001). The results of the logistic regression model analysis are summarized in Table 1. The signs of the β values, representing the coefficients and the constant in the equation, indicate the direction of each predictor's relationship to a damage outcome; thus, all the predictors, other than AT, were positively related to damage. X^2 Wald is related to the strength of each predictor. WQF and AT were the strongest predictors.

Table 1. Logistic Regression Table, Experiment 1. β values are the coefficients and constant of the model equation. SE β is the standard error of β. Wald $\boldsymbol{X^2}$ is a metric of the strength of each predictor. p< is the significance level of each predictor.

Predictor	**β**	**SE β**	**Wald X^2**	***p*<**
Constant	2.17	.31	6.96	.0001
WTAA	.00007	.000009	7.32	.0001
WQF	.11	.007	14.65	.0001
AT	-.00027	.00002	-14.07	.0001

The model fit the data quite well. One measure of fit is the *C* statistic, which assesses the proportion of all pairs of PIT events with different observed outcomes which the model predicts correctly. The *C* value of the model was .96, which is considered excellent as values greater than .80 are considered strong [17]. Thus, for 96% of all relevant pairs of events, the model correctly assigned a higher probability of damage to PIT events that resulted in damage than to events that did not result in damage.

2) Receiver-Operating Characteristic Analysis

Receiver-operating characteristic (ROC) analysis predicts how many damage events from the data were actually predicted by the logistic regression model [18]. Thus, each of the 1,999 PIT events was classified using the model and the results were then compared to the actual outcome for that event. In order to classify model outputs according to a binary outcome, such as damage vs. no-damage, a threshold value must be determined, with model outputs falling above that value classified as damage and those falling below the threshold classified as no-damage predictions. A ROC analysis determines the optimal threshold value for maximizing true positive classification and minimizing false positives. Figure 2 plots the

proportion of true positive and false positive classifications for each threshold. The optimal threshold is the one that maximizes true positive classifications and minimizes false positive classifications and thus corresponds to the upper left-hand point on the curve. The threshold for our model was determined to be .26. To classify a PIT event instance, the logit value output by the model equation is converted to a probability and then compared to the threshold; probabilities greater than the threshold predict an outcome of damage, while probabilities less than the threshold predict no damage.

ROC analysis also provides metrics to evaluate the classifications provided by the model. The area under the ROC curve (AUC) represents the probability that the model will rank a randomly selected positive instance (i.e., damage event) higher than a randomly selected negative instance (i.e., no damage event), and is thus similar to C [18]. Both AUC and C were equal to .96. Finally, ROC analysis provides an overall measure of fit, d´, which was equal to 2.65 for the current model. These results are indicative of a highly precise discrimination [19]. The rate of true positive classifications was high (87%), the rate of false positive classifications was low (6%). The results of the ROC analyses, as well as the C score, are displayed in the first row of Table 2.

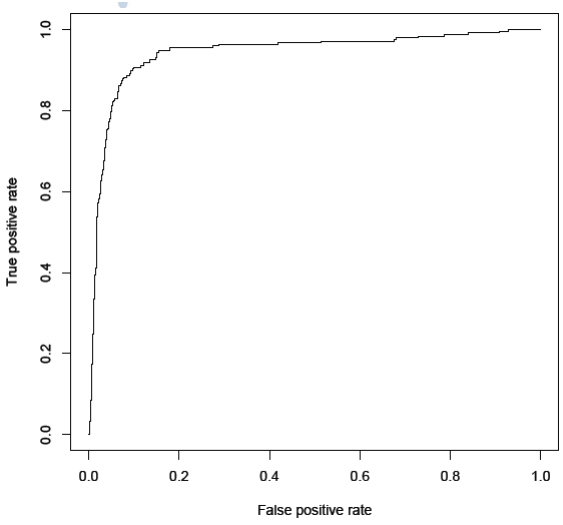

Figure 2. ROC curve for logistic regression model.

We chose to focus on a predictive model that concentrates on eye fixations rather than an alternate predictive model which more closely follows Cummings and Mitchell's fan-out model, using the variable WTQ in place of WQF. As can be seen in the second row of Table 2, the results of the WTQ model were not quite as good as the model using WQF, especially with regard to the true positive rate. For this reason, in addition to the benefits of the study eye movement described earlier, we explored the model described in Equation 4, which relied on the eye movement predictors WQF and WTAA, in the experiments that follow.

Table 2. Evaluation of Logistic Regressions Models using WQF and WTQ, Experiment 1

Model predictors	C	AUC	d´	True Positive (%)	False Positive (%)
WQF, WTAA, AT	.96	.96	2.65	87	6
WTQ, WTAA, AT	.94	.94	2.43	71	3

C. Discussion

The logistic regression model in Equation 4 was generated on the basis of the data in Experiment 1. The model, based on the fan-out model of supervisory control capacity of a human-multiple-robot control system, modeled operator overload on a per-event basis within an operator session. Operator overload was indicated by the occurrence of damage to a vehicle. The predictive value of the model's predictor variables may be understood in a comparable way to that of the corresponding fan-out variables. AT represents a task-constraint on the time available to solve a problem, in this case, the threat of damage to a vehicle. WQF assesses the operator's preoccupation with non-focal vehicles, which would clearly further constrain the time available to attending to the focal vehicle and possibly reduce the operator's visual attention to

the focal vehicle's status. Finally, WTAA is a measure of the operator's awareness of the focal vehicle's status, i.e., its need for attention. If an operator has noted the focal vehicle's status prior to attending to a non-focal vehicle, he/she is more likely to return to the focal vehicle after attending to the non-focal vehicle, thereby avoiding damage.

Multiple regression analytic methods, such as logistic regression, assume that the data points are independent, and this assumption is violated in the present model. Each participant contributes on average 57.1 PIT events to the data. As a practical matter, it would be difficult to gather data on 1,999 PIT events from that same number of participants. It would be possible in this case to use a mixed-model logistic regression model, but because those models separate fixed and random effects, they can be very difficult to use for prediction because random effects cannot be computed ahead of time for novel participants [20, 21]. The primary concern with not having independent data is that inferences may be incorrect and may not result in accurate generalizations to future datasets. We suspect that the data we collected has exchangeable random variables (future data will behave like past data, regardless of whether it is independent [22]), and the model we construct will generalize to future data sets. The strongest test of this model will occur throughout the rest of the paper where we show that the model does, in fact, generalize to other datasets and can even improve operator performance in real-time.

From a practical perspective, it would be desirable to have a predictive model that did not require the use of eye tracking equipment. We have shown that the predictor WTQ, which is not based on eye movements, is only moderately inferior to the eye-fixation predictor WQF. However we have been unable to find a satisfactory substitute for the eye-fixation predictor WTAA.

We developed a dynamic operator overload model from the data of Experiment 1 that was effective in predicting events in which UAVs incurred damage. Next, we will assess the robustness of the model, first, in Experiment 2, a replication of Experiment 1. Then we will assess the generalizability of the model to variants of the simulation task used in Experiments 1 and 2. People's strategies are often sensitive to small changes in task. Thus, assessments of the model's generalization to task variants will provide an assessment of the model's robustness.

IV. EXPERIMENT 2. REPLICATION FOR MODEL VALIDATION

In Experiment 2, we sought to determine how well the model generated from Experiment 1 data would generalize to the data obtained from an exact replication of Experiment 1. In this and all subsequent experiments to be reported here, the model being evaluated is defined by Equation 4 and is tested with the same threshold value, .26, determined by the ROC analysis of Experiment 1 data.

A. Method

The method in Experiment 2 was identical that in Experiment 1.

1) Participants

Forty-seven George Mason University undergraduate students participated for extra credit in a psychology course. No participant in any of the experiments reported here participated in more than one experiment. Six participants' data were not analyzed, because of a bug in the experimental software (4 cases) and running errors (2 cases). Thus, data from 41 participants, 13 male and 28 female, were analyzed. The mean video gaming experience of the participants was 1.8 (SD=.6).

B. Results and Discussion

Among the 41 participants in Experiment 2, there was a total of 2,679 PIT events, 273 (10.2%) of which ended in damage to the UAV.

The results of the ROC analysis are shown in Table 3, in row "2. Replication". As the table shows, the fit of the model to the data was excellent in Experiment 2 as it was in Experiment 1. The true positive rate was 81%, the false positive rate was 7%, $d'=2.41$, and AUC=.93. Thus, the model generalized well to the identical procedure to Experiment 1, providing evidence for the validity of the model. In the next two experiments, we assessed the

model's ability to predict errors under somewhat different procedures in order to further evaluate its robustness. Specifically, in Experiments 3 and 4, we evaluated the model under conditions expected to make it easier or harder, respectively, for the supervisor to prevent vehicle damage.

Table 3. Model evaluation across experiments.

Experiment	C	AUC	d´	True Pos. (%)	False Pos. (%)
1.Baseline	.96	.96	2.65	87	6
2.Repli-cation		.93	2.41	81	7
3. Easier		.95	2.56	86	7
4. Harder		.92	2.16	79	9

.

V. EXPERIMENT 3. MODEL GENERALIZATION TO AN EASIER TASK

Experiment 3 was designed to determine whether the model would generalize to an easier task, in which the vehicles had a higher level of autonomy. Specifically, in Experiment 3, participants were not required to deliver the payload using the payload panel. Instead, the system handled payload delivery automatically. As a result, engagement did not result in a task interruption as it did in the previous experiments, allowing the user to devote more cognitive resources to the problem of avoiding threats that cause vehicle damage.

A. Method

The method in Experiment 3 was identical to that in Experiments 1 and 2, except that the means of engagement did not involve an interruption involving the payload panel.

1) Participants

Thirty-three George Mason University undergraduate students, 8 male and 25 female, participated for extra credit in a psychology course. The mean video gaming experience of the participants was 1.7 (SD=.6).

2) Design and Procedure

The design and procedures in Experiment 1 were identical to those used in the previous experiments, except for the means of engagement. Engagement was complete when the participant right-clicked on a vehicle that had reached its target and selected the appropriate, engagement, menu item, as in the previous experiments. Unlike the previous experiments, there was no need to then deliver the payload by performing a visual identification subtask using the payload panel, and therefore no interruption of the main, map panel task. Following engagement, the vehicle was automatically re-assigned to a new target, as in the previous experiments.

B. Results and Discussion

There were 2,351 PIT events, of which 161 (7%) ended in damage. This contrasts with damage rates of 10.8% and 10.2% in Experiments 1 and 2, respectively. The damage rates in the experiments are summarized in Table 4. The higher incidence of damage in the previous experiments is likely attributable to the interruptions of the threat evasion task by the payload subtask, interfering with attention allocation[23, 24].

Table 4. Damage rate and payloads delivered in experiments.

Experiment	Damage rate (%)	Payloads delivered (mean)
1. Baseline	10.8	25.4
2. Replication	10.2	28.4
3. Easier	7.0	NA
4. Harder	13.9	27.0
5.a.No Cue condition	7.5	28.5
5.b.Cue condition	3.7	27.8

Despite displaying a lower damage rate, generalization of the logistic regression model to Experiment 3 was excellent. The results of the ROC analysis are given in Table 3, row "3. Easier." The true positive rate was 86%, the false positive rate was 7%, d′=2.56, and AUC=.95, all quite comparable to Experiments 1 and 2.

Thus, the model generalized well to an easier task. We next examined how well it generalized to a version of the task where it was more difficult to prevent damage.

VI. EXPERIMENT 4. MODEL GENERALIZATION TO A HARDER TASK

The task in Experiment 4 was made harder than in Experiment 1 by imposing a time constraint on engaging targets. Specifically, target engagement was only possible for 12 seconds following the arrival of a UAV on a target, after which the vehicle was automatically reassigned to a new target without delivering its payload. Thus, an added time constraint was imposed on participants in this condition, in addition to the time constraint for evading threats. In contrast to Experiment 3, where participants needed to pay *less* attention to target engagement than in the first two experiments, in the present experiment, participants needed to pay *more* attention to engagement. This demand was expected to divert cognitive resources from the task of avoiding threats so as to prevent vehicle damage, resulting in an increase in damage.

A. Method

Simulation and procedures were similar to Experiment 1, except for the time constraint on engagement.

1) Participants

Forty-seven George Mason University undergraduate students participated for extra credit in a psychology course. Six participants' data were eliminated, 5 due to an experimental bug and 1 due to an error in running the experiment. Thus, data from 41 participants, 14 male and 27 female, were analyzed. The mean video gaming experience of the participants was 1.9 (SD=.6).

2) Design and procedures

Procedures were identical to those used in Experiment 1. The simulation was identical to Experiment 1, except that if the participant failed to initiate payload delivery within 12 seconds after the UAV reached it, the UAV was reassigned to a new target without delivering its payload. As in Experiment 1, payload delivery was accomplished using the payload panel.

B. Results

Of the 2,676 PIT events in Experiment 4, 371 (13.9%) ended in damage (see Table 4). The higher rate of damage in comparison to Experiment 1 (10.6%) is consistent with Experiment 4 being a more difficult task. In contrast, the increased time constraint had little impact on performance of the other major subtask, payload delivery (see Table 4).

Despite the increase in damage, generalization of the damage prediction model based on Experiment 1 was very good, as the results in Table 3, row "4. Harder," show. The true positive rate was 79%, the false positive rate was 9%, d′=2.16, and AUC=.92. While very good, these results are not as strong as those for the easier task in Experiment 3.

In sum, Experiments 3 and 4 provided evidence for the robustness of the logistic regression dynamic operator overload model in

predicting damage under task variants resulting in either less or more damage than in Experiment 1, from which the model was derived. We next assessed whether the model's predictions could serve to help predict and prevent vehicle damage in real time.

VII. EXPERIMENT 5. MODEL-BASED CUES

While the demonstration of the performance of the dynamic operator overload model under various conditions argues for its robustness, it is not strictly a demonstration of the model's predictive power, since the model was generated and evaluated on data analyzed after the experiments were conducted. In an effort to support the claim that the model is truly predictive, we applied the model in real-time as a means to alert the operator to predicted damage. If providing cues alerting operators to impending damage results in a decreased rate of vehicle damage, it will provide evidence of the predictiveness of the model. This is what we attempted in the next experiment.

A. Method

Simulation and measures were identical to those used in Experiment 1 in the control, No Cue, condition and in the Cue condition; the method was identical to Experiment 1 except for the provision of adaptive cues. In the Cue condition, the logistic regression model was used in real time to predict whether a vehicle was likely to hit a threat area and, if it was, to alert the participant of the danger by flashing the relevant threat. The model re-assessed the status of each PIT every 500 ms, when the simulation updated itself.

1) Participants

Forty-three George Mason University undergraduate students participated for extra credit in a psychology course. Participants were assigned randomly to the No Cue control condition or to the Cue condition. Twenty-two participants, 10 male and 12 female, were in the No Cue condition and 21 participants, 9 male and 12 female, were in the Cue condition. The mean video gaming experience of the participants was 1.9 (SD=.6) in the No Cue condition and 1.9 (SD=.6) in the Cue condition.

2) Design and procedures

Procedures were identical to those used in Experiment 1, except that in the Cue condition the dynamic operator overload model (see Equation 4) was used as a basis for alerting the participant of an impending UAV encounter with a threat. The damage likelihood of each UAV on a path intersect threat course was computed every 500 ms using the dynamic operator overload model, and when the likelihood exceeded the threshold value derived from the ROC analysis in Experiment 1, the relevant threat was highlighted by turning blue (from yellow) and blinking to alert the user of the approaching threat. The blinking threat thus served as a predictive cue, alerting the user to the threat to which they needed to divert their attention when the system determined the operator was overloaded.

B. Results

Of the 1,448 PIT events in the No Cue condition, 108 (7.5%) ended in damage. In contrast to this, of the 1,396 PIT events in the Cue condition, only 52 (3.7%) of them ended in damage (See Table 4). Curiously, the No Cue condition witnessed less damage relative to Experiments 1 and 2, which had comparable procedures. We attribute this anomaly to random variation. In any case, the comparison of the two conditions in Experiment 5 showed that alerting the user to predicted damage via cues reduced the rate of damage by more than half. In addition, in all PIT events ending in damage in the Cue condition the cue appeared. That is, there were no cases of damage on PIT events where the cue failed to appear, i.e., no false negative errors. Thus, the cue was a strong predictor of damage and the cue was effective in preventing damage, as predicted by the dynamic operator overload model.

This suggests that the damage instances that occurred despite the cue's appearance had a different cause from damage instances predicted by the model. Indeed, as Table 3 shows, the model predicted damage very well in the No Cue condition, but not in the Cue condition of Experiment 5. We believe the remaining instances of damage, not prevented by the cue, were due to *concurrent urgent* PIT events, i.e. non-focal PIT events that triggered cues on the

basis of the damage prediction model (i.e., urgent) and that overlapped in time with the focal event (i.e., concurrent). In the Cue condition, there was a mean of .84 (SD=1.09) concurrent urgent PIT events. For urgent PIT focal events not ending in damage, the rate of concurrent urgent PIT events was similar to this baseline (M=.96, SD=1.13). However, for urgent PIT focal events ending in damage, there were about twice as many concurrent urgent PIT events as the baseline (M=1.98, SD=1.33). Thus, damage incurred despite the cue was associated with competition between multiple concurrent urgent PIT events. In such situations, damage could be avoided only if the focal event was the one first selected by the operator to handle. If not given priority by the operator, such urgent PIT events ended in damage. Thus, damage that occurred despite the model-generated cue was likely due to task overload that exceeded the operator's capacity.

We believe the cue served primarily to encourage attention to the threat, rather than to support cognition of the threat. The RESCHU user interface clearly represents the UAVs' trajectories towards their respective targets graphically by lines. Thus an upcoming threat could be recognized on a perceptual basis by the visible trajectory's traversal of a threat area. The cue likely drew the users' attention to threats they had not noticed or had forgotten.

In sum, the results of Experiment 5 provided further support for the dynamic operator overload model. The model served as a basis for real-time cues to alert the operator to impending vehicle damage. The cues reduced the damage rate by about half and never were presented in events where there was no damage. Cases where damage occurred despite the cue were characterized by simultaneous cues for more than one threat, indicating that the overloading was too great for damage from all co-occurring threats to be avoided. Thus, the dynamic operator overload model appears to be a good predictor of supervisor overload.

VIII. Fan-out and Performance Prediction

Fan-out values for each experiment are displayed in Table 5, computed using Equation 3, which we will refer to as "AT fan-out", based either on all PIT events and based only on PIT events where no damage occurred. Note that since fan-out here is computed only based on PIT events, not on payload events, it does not provide a complete assessment of the demands of the respective tasks. The fan-out values displayed in Table 5 suggest that operators in our experiments were often required to supervise somewhat more vehicles (i.e. five vehicles) than recommended by the fan-out model. The main exception was the easier task in Experiment 3. In that experiment, the absence of a competing task, payload delivery, reduced the amount of time devoted to competing tasks (i.e., WTQ) during PIT events.

Table 5. Fan-out values for all experiments. Operators controlled 5 UAVs in all experiments.

Experiment	AT Fan-out (all)	AT Fan-out (no damage)	NT Fan-out (no damage)
1. Baseline	3.8	4.6	4.8
2. Replication	4.4	5.3	4.9
3. Easier	5.9	6.7	5.1
4. Harder	3.5	4.1	3.9
5.a No Cue condition	4.5	5.3	5.1
5.b Cue condition	4.0	4.2	4.1

We wished to determine whether AT fan-out based on available time, as in Equation 3, is similar to fan-out based on neglect time, as in Cummings and Mitchell, Equation 2. We computed NT fan-out as follows:

$$\text{NT fan-out} = (\text{NT}/\text{PIT.duration}) + 1 \quad \textbf{(5)}$$

where NT is time outside of PIT and of payload delivery events and PIT.duration is the duration of PIT events. PIT.duration served as a substitute for the expression in the denominator of Equation 2, involving WTAA, WTQ, and IT,

since the entire PIT event consists of some combination of these. During a PIT event, there is always an object that needs attention, namely the focal object. Thus every moment in the event, the participant is either working on/attending to the focal object or not working on it. In the former case, the time represents interaction time (IT), in the latter case it represents wait time (WT). Wait time is either WTQ or WTAA depending on whether the participant is attending to non-focal objects or not, respectively. Thus, the entire PIT event represents some combination of the denominator variables in Equation 2. Further, no time outside of the PIT event contributes to those variables, with the exception of time spent on the focal vehicle's payload delivery, but that interval is not included in NT.

As Table 5 shows, fan-out values computed based on NT were similar to those based on AT. The principal exception was in Experiment 3 again, where NT fan-out was lower than AT fan-out. The similarity between the two measures of fan-out supports our interpretation of our logistic regression model as a dynamic version of the Cummings and Mitchell fan-out model.

A comparison among the first 4 experiments demonstrates that relative task difficulty was reflected similarly in fan-out and damage rate. However, the intervention of providing cues in Experiment 5 did not improve fan-out, even though it radically reduced the damage rate. This suggests an important difference between fan-out and performance prediction. Fan-out is concerned with having enough time to perform a task, whereas prediction is primarily concerned with what users do with the available time. The cue does not change the amount of time required to perform subtasks but does alert users to direct their efforts to a particular subtask requiring immediate attention. Time intervals where the operator lacked attentional awareness of one problem were not moments of idleness, as the current task is a highly dynamic, time-pressured task in which the operator is continually active. In moments in which the operator has lost awareness of one problem, they are likely engaged on another problem. As a result, lost awareness may result in poorer performance without increasing the overall time required for the task.

Another important difference between damage prediction and fan-out is suggested by a predictive model based on a single variable, i.e. the time remaining to work on the focal threat problem after consuming time on WTAA delay and on working on other, non-focal objects. This duration thus represents potential interaction time:

$$\text{potential-IT} = AT - (WTAA + WTQ_1) \quad \textbf{(6)}$$

where WTQ_1 includes both time spent acting on non-focal objects and time spent fixating such objects and where both of those durations are calculated so as not to overlap with the WTAA interval, i.e. the initial duration of the PIT before the focal threat is first fixated. As Table 6 shows, the single-variable model performs as well as the dynamic operator overload model, defined by Equation 4. Further, the single-variable model adheres more closely to the principle underlying the fan-out equation, as it is based solely on an estimate of the time remaining to work on the focal problem after deducting all wait times from the available time.

Table 6. Comparison of the dynamic operator overload model and a model based on Potential-IT, Experiment 1 data

Model	C	AUC	d´	True Posit-ive	False Posit ive
Dynamic Operator overload Model	.96	.96	2.65	87%	6%
Potential-IT Model	.97	.97	2.83	82%	3%

However, this model can be shown to be much less predictive than the original model in terms of the delay between prediction and event predicted. In the Cue experiment (Experiment 5), the dynamic operator overload model produced a warning signal after only 20% of the available time (AT) had elapsed, on average. In contrast, solving the single-variable model's logistic regression equation for the value required to exceed the model's threshold shows

the model would not produce a warning signal until 90% of the available time had elapsed. That is too late to be useful to the operator. Fan-outs based on Equation 3 under procedures similar to the Experiment 5 control condition (i.e., Experiments 1 and 2) are generally between 4 and 5 (see Table 5). Thus, approximately 20-25% of available time is required to take care of each vehicle. Warnings provided by the potential-IT model, when only 10% of available time remains, are clearly insufficient, while the warnings provided by the dynamic operator overload model are more than adequate, allowing 80% of the available time to take care of the vehicle needing attention.

These observations highlight one key difference between fan-out and predictive models. Fan-out models analyze performance on a task globally and so are not aimed at within-task prediction, beyond a global prediction of how many vehicles an operator will be able to supervise in a given task. The dynamic operator overload model, in contrast, is useful for prediction of performance of events *within* a task session.

However, the comparison of the two logistic regression models demonstrates that even logistic regression does not always provide useful predictions of real-time performance. Logistic regression makes no distinction among the relative position of data points in a time sequence and so requires theory-based selection of possible predictors by the researcher in order to contribute to prediction. Obviously, the sooner a prediction can be made the better; indeed, it would appear almost a tautology to say that a more timely prediction is more predictive than a delayed prediction, assuming both are equally accurate. The variable potential-IT diminishes progressively from the start to the end of the PIT event. In contrast, in the dynamic operator overload model, the value of one predictor, Available Time, is known at the start of the PIT event. A second predictor, WTAA is an interval that begins at the start of the event and that usually ends well before the event is finished. Only the third predictor, WQF (or WTQ), grows progressively throughout the course of the event. As a result, the dynamic operator overload model provides a better basis for an alert system than the potential-IT (Equation 6) model.

In addition, a predictor to be useful must be theoretically meaningful. This is illustrated by the potential-IT model, which is based only on time remaining to perform a task. It's trivial that a participant who never performs an action will run out of sufficient time to do so shortly before the deadline. To paraphrase a well-known saying, it's always darkest before nightfall. It is more useful from a theoretical and practical perspective to know that the operator's failure to notice a problem and the operator's preoccupation with other objects and activities predict the failure to correct the problem. More generally, high scores on typical criteria to assess logistic regression models (C, d′, etc.) are not sufficient to guarantee that a model is theoretically or practically useful.

IX. CONCLUSION

Dynamic operator overload was introduced as an assessment of the overall human-robot system in supervisory control applications. The fan-out model of Cummings et al., for instance, was designed to estimate the number of UAVs a single operator can supervise. Its estimate is based on time intervals, including the length of time a vehicle may be ignored before its performance degrades below a specified threshold (NT), the time required to bring a vehicle's performance back up above the threshold (IT), and delays between those two intervals due to loss of attentional awareness (WTAA) and due to time spent on higher-priority tasks (WTQ).

We explored the relationship between *system-focused* fan-out, on which Cummings et al. and others have focused, and *dynamic operator overload*, which varies over the course of operator-system interaction. Even when an operator is required to supervise no more vehicles than dictated by the system-focused fan-out model, there may be moments when dynamic task demands converge to overload the operator, resulting in error. It is the goal of a dynamic operator overload model to predict such situations of transitory overload.

In the current research, the dynamic operator overload model was developed to predict the

quality of ongoing performance of novice operators engaged in a simulation task in which they supervised five UAVs, attempting to keep them from incurring damage by traversing threat areas while directing the vehicles to deliver payloads on assigned targets. We developed logistic regression models to predict vehicle damage based on ongoing operator behaviors and attention as assessed by operator eye movements. Our models took the variables that figure in system-focused fan-out models as their starting point.

The fan-out variables WTAA and WTQ, together with the variable *available time* (AT) (substituted for NT for task-specific reasons), yielded a model that was highly predictive of damage occurrences in path-intersect threat (PIT) events. The time required to perform an action to avoid a threat (i.e., IT) offered insufficient variability to figure within the model for our task, but might contribute to models in situations where it exhibits greater variability.

More predictive still was a model that substituted number of fixations on non-focal objects (WFQ) in place of time spent acting on non-focal objects (WTQ). This model's parameters were generated from the data in Experiment 1. The model was then replicated in Experiment 2, generalized to an easier task and a harder task in Experiments 3 and 4, respectively, demonstrating the model's robustness. The model was then applied in Experiment 5 where the model initiated cues that alerted the user to impending damage. The superiority of this model, with the WFQ variable, over the model using WTQ may be due to the fact that fixations are a more comprehensive measure than manual actions and are more sensitive to individual differences in visual and/or core processing speed. Fixations on objects occur during manual actions on those objects (i.e., as in WTQ) as well as during cognitive activity in the absence of overt action, such as scanning and decision making.

The research also pointed to the importance of attention in predicting performance. Two of the three variables in the dynamic operator overload model, WTAA and WFQ, were based on eye fixation data. WTAA, the time it took for the operator to first fixate on the relevant threat of a path-intersects threat event, is clearly related to attention. WFQ, the number of fixations on non-focal objects may reflect attention in part. The success of the alert cue in Experiment 5 in greatly reducing damage rate likewise suggests the importance of attention to task performance.

The success of the model-based cues in Experiment 5 also provided evidence of the ability of the dynamic operator overload model to predict damage in real-time. The cues reduced the rate of damage by about half. What is more, no damage occurred in the absence of a cue. The success of the cues also points to the potential practical application of the dynamic operator overload model.

The research also shed light on considerations involved in developing a model that provides *timely* feedback, as far in advance as possible. We demonstrated that a model that is a highly discriminating classifier may not be able to make a decision until late in an event, at which point a warning may come too late to be useful. In contrast, our model, when used as the basis for user cues, was able to alert the user of a threat after only 23% of the event had passed. We argued that the timeliness of the cue was due to the model's reliance on factors most of which were available at, or soon after, the start of the event.

The main conclusion of this research is that a system-focused fan-out model may be adapted to produce a dynamic operator overload model. Both models rely on the operator's allocation of available time to competing subtasks. Whereas system-focused fan-out is a global assessment of a task, the dynamic operator overload model allows for variations in both available time and the number of competing time demands during the course of a task, the focus of dynamic operator overload. Some of this variation is imposed by the environment, outside of the operator's control, whereas other variations in time allocation are due to the operator's attentional awareness and the operator's decision and planning skills. Both types of variation contributed to the dynamic operator overload model presented here.

ACKNOWLEDGMENTS

The authors thank Missy Cummings and the Human Automation Lab at MIT for providing us with RESCHU. The authors also thank Joo Park and Giorgia Picci for providing assistance in conducting the experiments.

REFERENCES

[1] J. W. Crandall, M. A. Goodrich, D. R. Olsen, and C. W. Nielsen, "Validating human-robot interaction schemes in multitasking environments," *IEEE Systems, Man, and Cybernetics,* vol. 35, pp. 438-449, 2005.

[2] M. L. Cummings and P. J. Mitchell, "Predicting controller capacity in supervisory control of multiple UAVs," *IEEE Systems, Man, and Cybernetcis, Part A: Systems and Humans,* vol. 38, pp. 451-460, 2008.

[3] J. W. Crandall, M. L. Cummings, and C. E. Nehme, "A predictive model for human-unmanned vehicle systems," *J. Aerosp. Comput. Inf. Commun.,* vol. 6, pp. 585-603, 2009.

[4] J. W. Crandall, M. L. Cummings, M. Della Penna, and P. M. A. de Jong, "Computing the effects of operator attention allocation in human control of multiple robots," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 41, pp. 385-397, 2011.

[5] K. Rayner and R. K. Morris, "Do eye-movements reflect higher order processes in reading?," in *From eye to mind. Information acquisition in perception, search, and reading*, R. Groner, D. d'Ydewalle, and R. Parham, Eds. Amsterdam: North-Holland, 1990, pp. 191-204.

[6] K. Rayner, "Eye movements in reading and information processing: 20 years of research," *Psychological Bulletin,* vol. 85, pp. 618-660, 1998.

[7] M. A. Just and P. A. Carpenter, "Eye fixations and cognitive processes " *Cognitive Psychology,* vol. 8, pp. 441-480, 1976.

[8] I. D. Brown, "A Review of the 'Looked-But-Failed-To-See' Accident Causation Factor; Road Safety Research Report No. 60," Department for Transport, UK, 2002.

[9] J. M. Henderson and A. Hollingworth, " The role of fixation position in detecting scene changes across saccades.," *Psychological Science,* vol. 10, pp. 438 – 443, 1999.

[10] M. S. Peterson, A. F. Krame , and D. E. Irwin, "Covert shifts of attention precede involuntary eye movements.," *Perception and Psychophysics* vol. 66, pp. 398 – 405, 2004.

[11] K. Rayner, G. W. McConkie, and S. Ehrlich, "Eye movements and integrating information across fixations.," *Journal of Experimental Psychology: Human Perception and Performance,* vol. 4, pp. 529 – 544, 1978.

[12] R. M. Ratwani and J. G. Trafton, "A generalized model for predicting postcompletion errors," *Topics in Cognitive Science,* vol. 2, pp. 154-167, 2010.

[13] R. M. Ratwani, J. M. McCurry, and J. G. Trafton, "Single operator, multiple robots: An eye movement based theoretic model of operator situation awareness.," in *Proceedings of the conference on HRI*: IEEE, 2010, pp. 235-242.

[14] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The Journal of Educational Research,* vol. 96, pp. 3-14, 2002.

[15] Y. Boussemart and M. L. Cummings, "Behavioral recognition and prediction of an operator supervising multiple heterogeneous unmanned vehicles," *Humans operating unmanned systems,* 2008.

[16] C. Nehme, "Modeling human supervisory control in heterogeneous unmanned vehicle systems," in *Department of Aeronautics and Astronautics* Cambridge, MA: MIT, 2009.

[17] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 2nd ed. New York, NY: John Wiley & Sons, 2000.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, pp. 861-874, 2006.

[19] J. A. Swets, "Signal detection theory and ROC analysis in psychology and diagnostics: Collected Papers.," Mahwah, New Jersey: Lawrence Erlbaum Associates, 1996.

[20] D. Bates, "lme4: Mixed-effects modeling with R," in *http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf*: http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf, 2010.

[21] D. Bates and M. Maechler, " lme4: Linear mixed-effects models using S4 classes," in *http://CRAN.R-project.org/package=lme4*, 2009.

[22] G. J. Szekely and J. G. Kerns, "De Finetti's theorem for abstract finite exchangeable sequences," *Journal of Theoretical Probability,* vol. 19, pp. 589–608, 2006.

[23] E. M. Altmann and J. G. Trafton, "Memory for goals: An activation-based model," *Cognitive Science,* vol. 26, pp. 39-83, 2002.

[24] J. G. Trafton, E. M. Altmann, and R. M. Ratwani, "A memory for goals model of sequence errors," *Cognitive Systems Research,* vol. 12, pp. 134-143, 2011.

List of Figures

Footnotes

Manuscript received …

Leonard A. Breslow is a cognitive scientist at the Naval Research Laboratory, Code 5515, Washington DC 20375; phone: 301-602-3585; email: len.breslow@nrl.navy.mil

Daniel Gartenberg is a Ph.D. student at George Mason University, Fairfax VA; email: dgartenb@masonlive.gmu.edu

J. Malcolm McCurry is a research scientist at ITT Exelis, McLean, VA; email: malcolm.mccurry.ctr@nrl.navy.mil.

J. Gregory Trafton is a cognitive scientist at the Naval Research Laboratory, Washington DC; email: greg.trafton@nrl.navy.mil.

This work was supported in part by the Office of Naval Research under funding document N0001409WX20173 and N0001410WX30037 to JGT. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Navy.